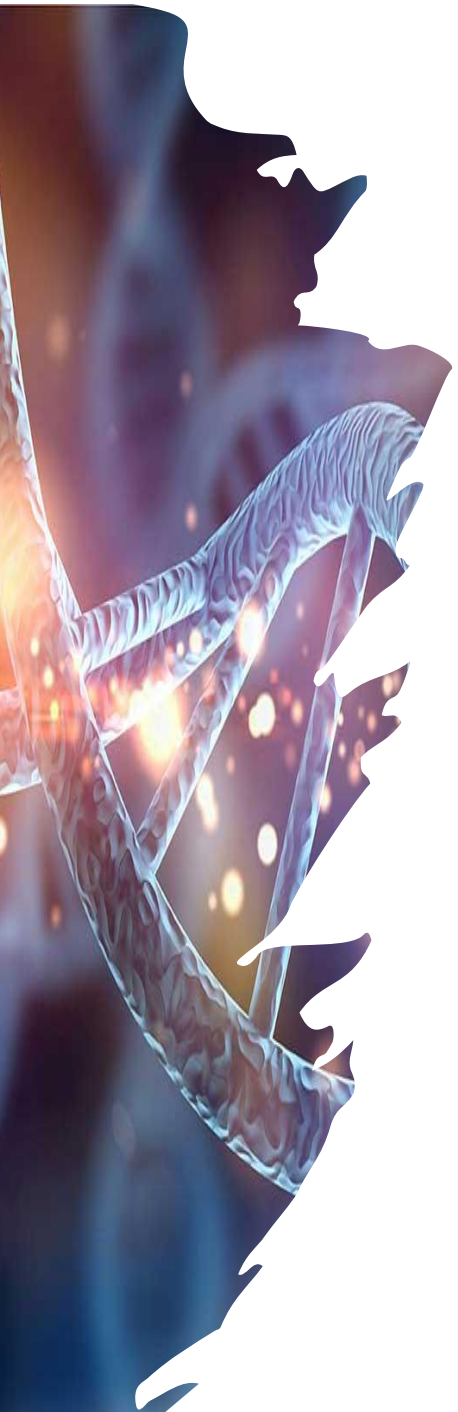
Cancer Gene Detection

Use Case Scenario Advance Analytics Application

Benazir de la Rosa

Presentation Outline

- **Problem Definition**
- Data Description
- Descriptive Analytics
- Natural Language Processing
- Graph Network Analysis
- Machine Learning Algorithm
- Results and Next Steps

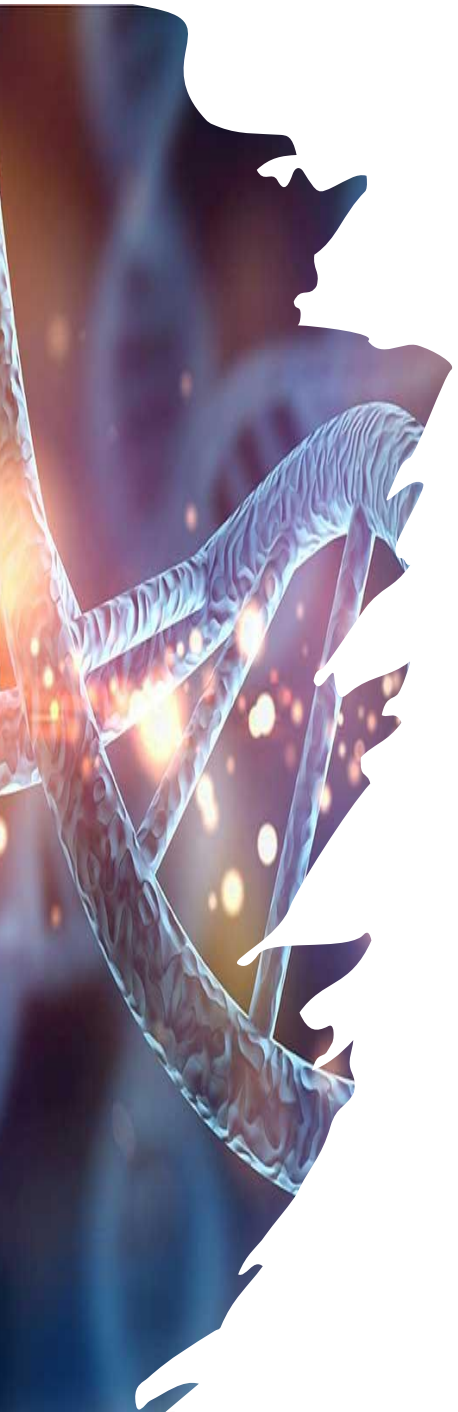


Problem Definition

- Cancer is a human disease characterize by the uncontrolled growth and division of abnormal cells.
- Diagnostic cancer within early stages is a source to saving human lives. Thus, the focus of this project is to identify proteins linked to malign cancer.
- We want to state that by predicting high activity on protein interactions which are highly likely to be malign cancer.
- The project will perform advance analytics over the data to reach out cancer detection by using descriptive analytics, natural language processing, machine learning and data processing techniques.

Presentation Outline

- Problem Definition
- **Data Description**
- Descriptive Analytics
- Natural Language Processing
- Graph Network Analysis
- Machine Learning Algorithm
- Results and Next Steps

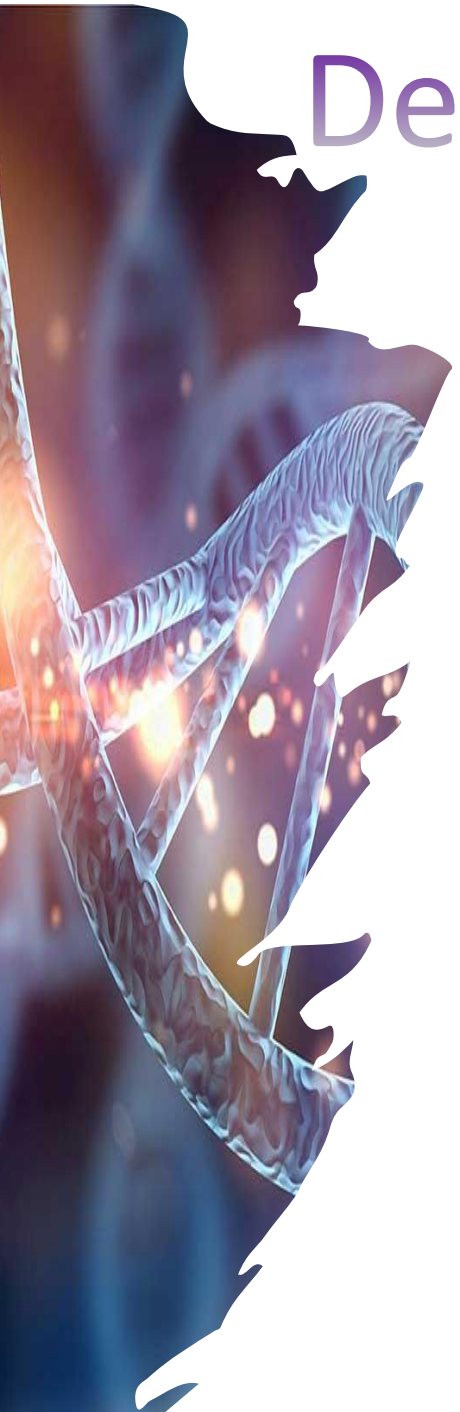


Data Description

- The data was extracted from https://string-db.org. The website contains protein interactions from different kind of networks. I took the ones related to humans.
- The first dataset “9606.protein.info.v11.0” (refer as “Protein Description”) contains the next attributes, its main purpose is to describe each protein:
  - Protein_external_id: ID protein
  - Preferred_name: protein nickname
  - Protein_size: protein size
  - Annotation: protein description
- The second dataset “9606.protein.links.detailed.v11.0.txt” (“refer as Protein Network”) contains the next attributes, its main purpose is to describe each protein-protein interaction :
  - Protein1: In the protein-protein interaction protein1
  - Protein2: In the protein-protein interaction protein2
  - Neighborhood: Location
  - Fusion: Property of the interaction
  - Coocurrence: Property of the interaction
  - Coexpression: Property of the interaction
  - Experimental: Lab label
  - Database: DB location
  - Textmining: NLP data
  - Combined_Score: Lab label

Presentation Outline

- Problem Definition
- Data Description
- **Descriptive Analytics**
- Natural Language Processing
- Graph Network Analysis
- Machine Learning Algorithm
- Results and Next Steps

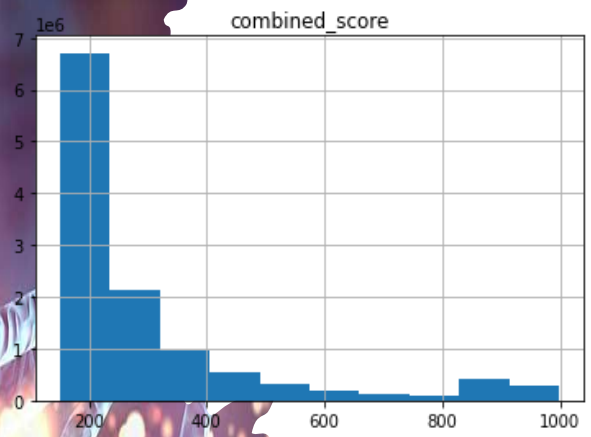


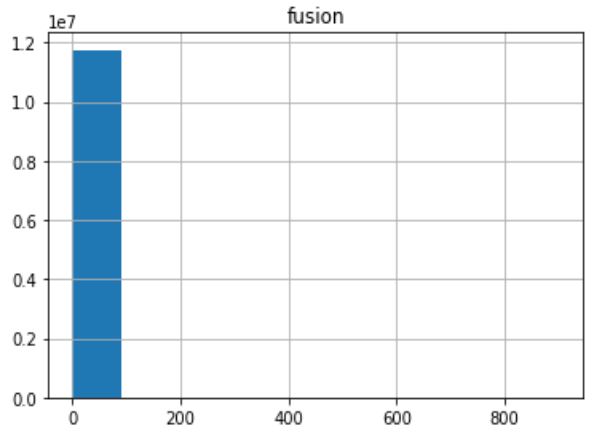
Descriptive Analytics: Direct Numbers

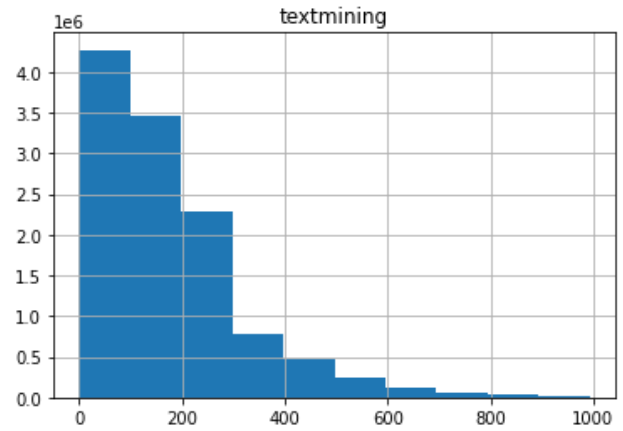
- Protein Description Dataset:
  - Describes 19566 distinct proteins.
  - It was found that 199 description contained the “cancer” noun.
  - 132 descriptions with “cancer” noun turned out to related to malign cancer.
  - 67 descriptions with “cancer” noun turned out to related to positive cancer, meaning they help to prevent it.
  - In total only **0.6% proteins represented malign cancer.**

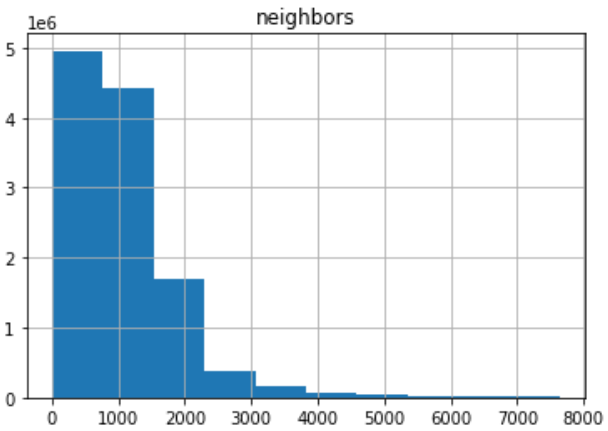
- Protein Network Dataset:
  - It contains **11759453 connections between proteins**.
  - No label related cancer assign.

Descriptive Analytics: Protein Network





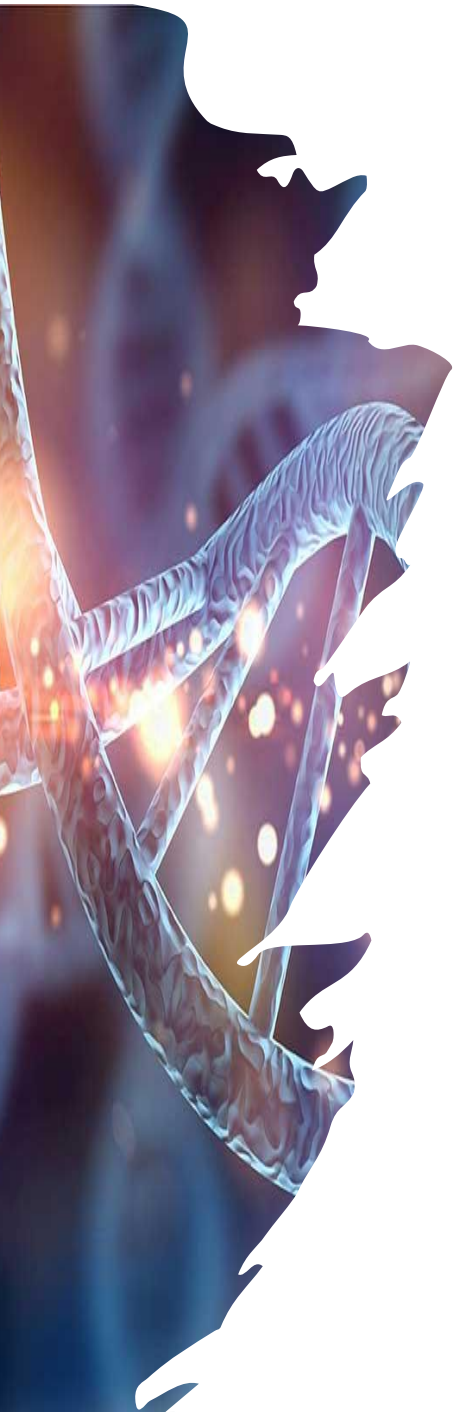




- The distributions of the most significant features on the Protein Network dataset look as follows:
  - The most features had almost no distribution present such as the fusion example.
  - The combined_score and textmining feature where the ones with a nice skewed to the left distribution.
  - The neighbors attribute was brough as feature engineering from analyzing the graph network. It has a nice skewed to the left distribution.
  - Overall, the features in the protein network have a median bigger than the mean.

Presentation Outline

- Problem Definition
- Data Description
- Descriptive Analytics
- **Natural Language Processing**
- Graph Network Analysis
- Machine Learning Algorithm
- Results and Next Steps

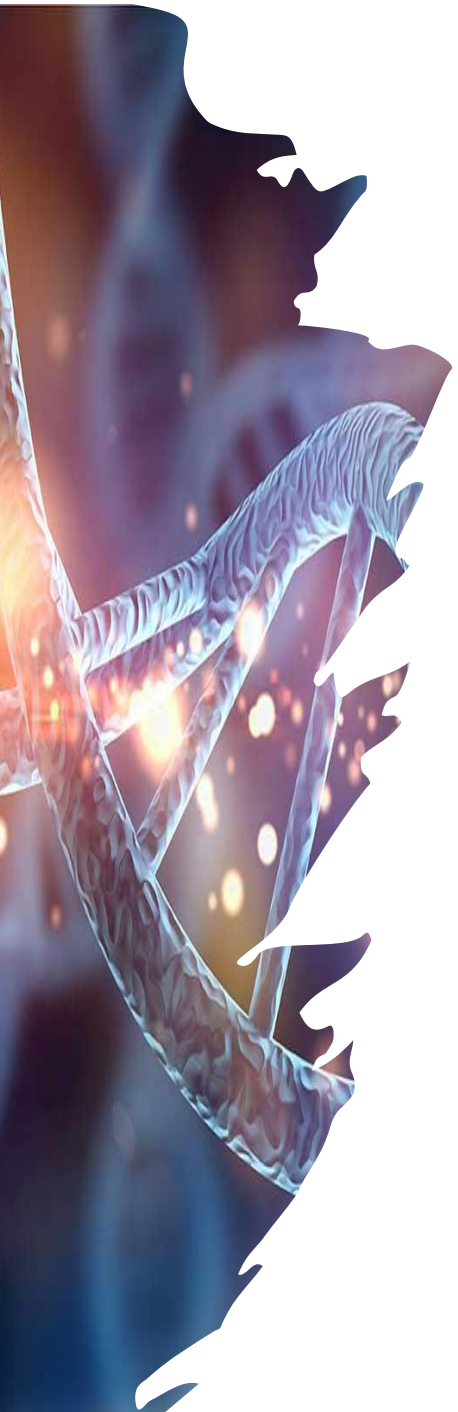


Natural Language Processing

- I use natural language processing on the protein description dataset to try to infer, the cancer label out of the description.

- The intuition behind the idea was that given that my description has a negative sentiment and the cancer word it is more likely that it will be represent maligne cancer.

- Overall, the NLP pipeline has:
  - Text cleaning, numbers, dots and line creation.
  - Tokenization
  - Stopwords erasing
  - Lemmatization, because I believe that a stem will cut too much the word making it lose the sense
  - Sentiment analysis with the NLTK *SentimentIntensityAnalyzer()*
  - A label to identify whether the description contains the word cancer or not

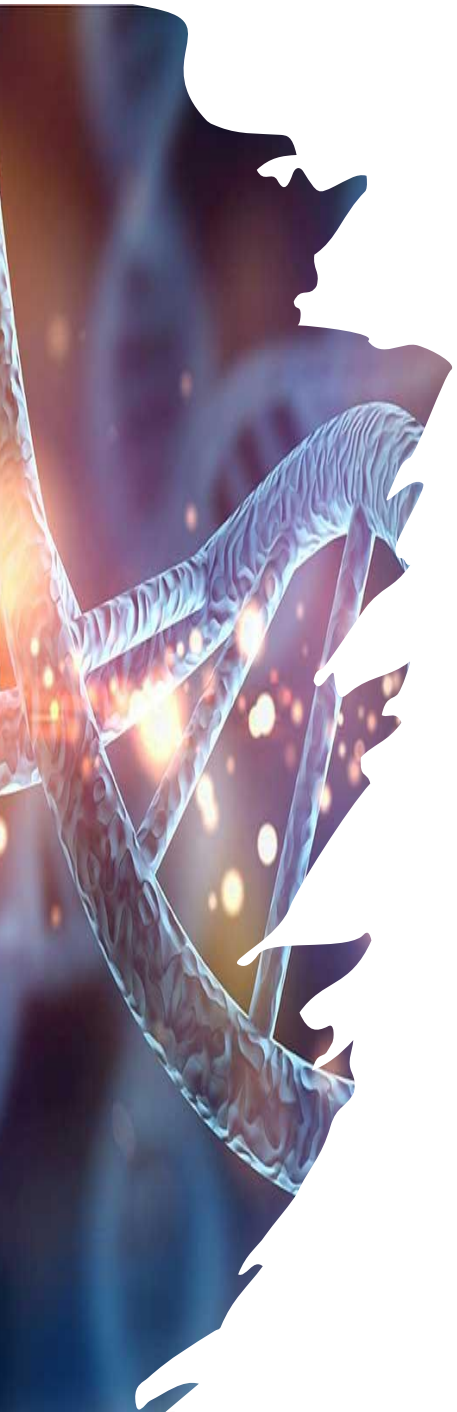


Natural Language Processing

- The NLP results, generating the cancer label:
- Overall to detect cancer out of the description by applying sentiment analysis gave a good result with the negative descriptions. From the 20 cases detected with malign cancer using sentiment analysis only 1 case was a False positive.
- The chances of classifying wrong a negative descriptions are %5
- The intuition behind a *"negative description + cancer noun = malign cancer"* turn out to work quite well.
- However, there was a big problem while working with the positive descriptions. From the 179 cases related to proteins which protect from cancer, 111 cases turn out to be wrong classify. The problem took place because some verbs such as "promotes" cancer, are not considered to be negative.
- The chances of classifying wrong a positive description are 62%
- Therefore, the intuition behind a *"positive description + cancer noun = no malign cancer"* turn out to not work well.
- When comparing the sentiment analysis with the ground truth data. Annotated by me. The sentiment analysis labeling classification based on the description sentiment turn out to have a 0.25 accuracy.
- Thus, I use the cancer detected data by me. However, it is quite positive that **it works to detect malign cancer with negative descriptions** most of the time.

Presentation Outline

- Problem Definition
- Data Description
- Descriptive Analytics
- Natural Language Processing
- **Graph Network Analysis**
- Machine Learning Algorithm
- Results and Next Steps

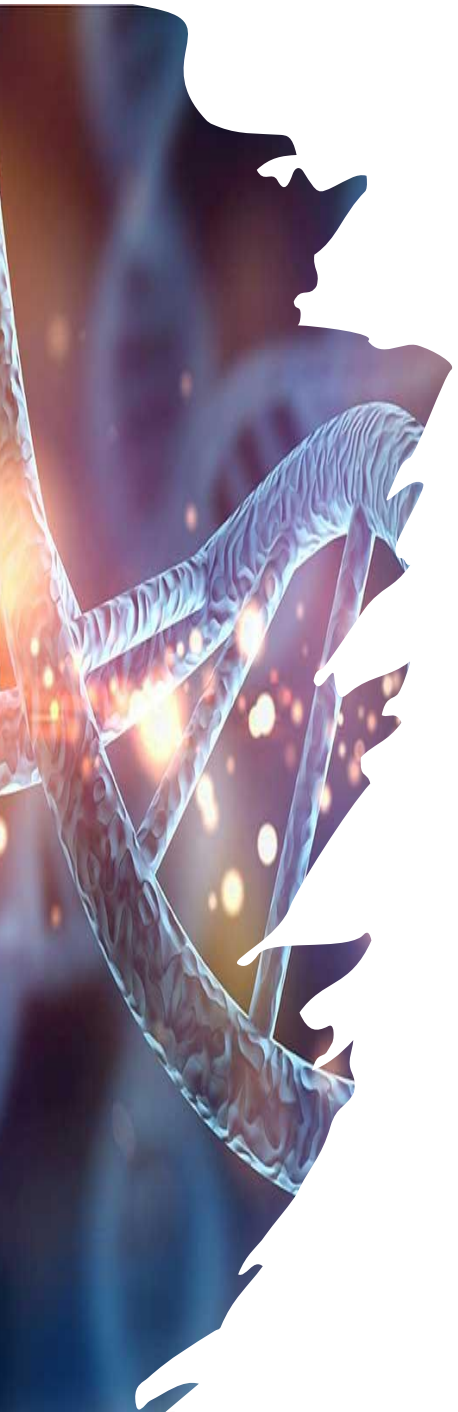


Graph Network Analysis

- A directed graph was created.
- The edges turned out to be the **same for in-degress and out-degrees**.
- The Protein Network behaves as ***Undirect Graph***.
- **Compute neighbors** per node (Protein ID)
- The more connection I do have, the more likely it will turn out to be a malign cancer protein. (*Intuition*)

Presentation Outline

- Problem Definition
- Data Description
- Descriptive Analytics
- Natural Language Processing
- Graph Network Analysis
- **Machine Learning Algorithm**
- Results and Next Steps



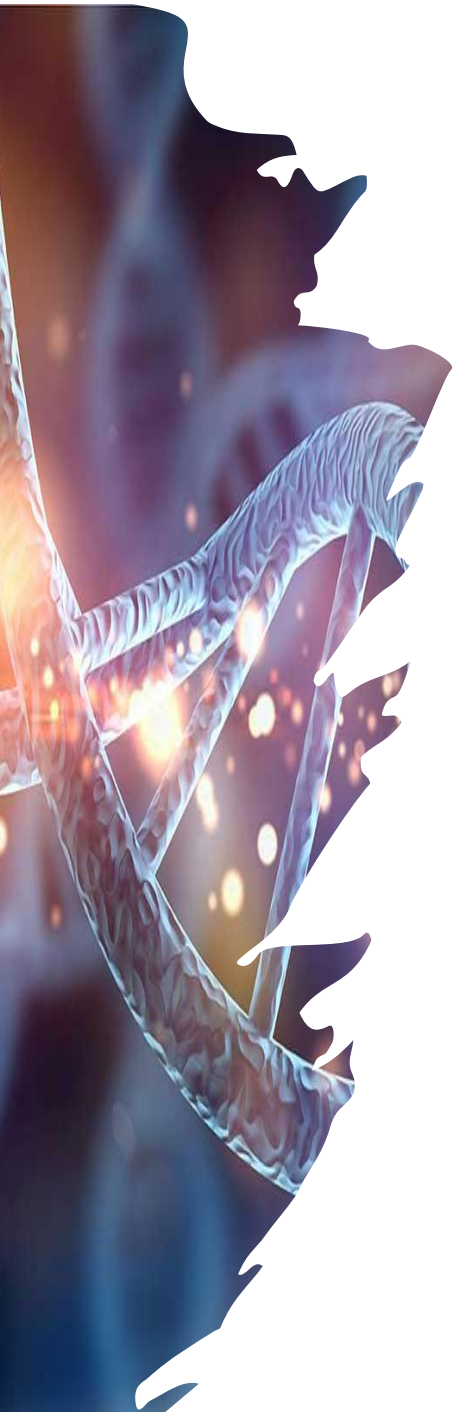
Machine Learning Algorithm

Feature Engineering

- Add Neighbors to the Protein Network dataset.
- In the features of combined_score, textmining and neighbors. Subtract IQR and add them as additional 3 columns.
- Balance the data, because only ~2% was malign cancer.
- Data encode and normalization.
- PCA.

Presentation Outline

- Problem Definition
- Data Description
- Descriptive Analytics
- Natural Language Processing
- Graph Network Analysis
- Machine Learning Algorithm
- **Results and Next Steps**



Results

The results from the baseline are as follows:

- The selected baseline algorithm to predict the cancer was a Support Vector Machine Algorithm.
- The algorithm resulted in a ~85% accuracy. It was created considering a cross validation evaluation process with a K5 fold.

Next Steps

- Improve feature engineering.
- Test more algorithms.
- Add more information from different sources.
- Include more cancer cases.

Thank you!

Benazir de la Rosa. Data Science Consultant.